CORRESPONDENCE ANALYSIS APPLICATION IN CLASS COMPARISON STUDIES

PhUSE London 2014

Edyta Winciorek
FROM GENE EXPRESSION TO EXPRESSION CARTOGRAPHY

1. MICROARRAY GENE EXPRESSION PROFILE PRELIMINARIES

2. TYPES OF DNA MICROARRAY EXPERIMENTS

3. EXAMPLE - CORRESPONDENCE ANALYSIS APPLICATION IN CLASS COMPARISON STUDIES

4. CONCLUSION
MICROARRAY GENE EXPRESSION PROFILE PRELIMINARIES
Hybridization is the process of joining two complimentary strands of DNA or one of DNA and RNA to form a double-stranded molecule.

DNA double-stranded structure is formed through the hybridization of these complementary single stranded chains.

Given these characteristics, it is possible to detect a target gene by hybridizing it with DNA that has a complementary sequence.

DNA ARRAY

glass slide where **single-stranded** DNA (called **probe**) with various sequences are printed on the surface of the substrate in a localized features that are arranged in a grid-like pattern.

Reverse transcribed mRNA to given unique cDNA population

Embed population onto specially coated glass slides

The slides are coated with positively charged polysine. DNA is negatively charged, so that cDNA stick to the slide through an ionic interaction.

The cDNA from ill tissue (e.g. tumor) are labelled with red fluorescent tag Cyanine 5 (Cy5).

The cDNA from normal (reference) tissue are labelled with green tag Cyanine 3 (Cy3).

Tagged cDNA arrays are incubated, which bound the matching genes printed on the arrays.

If the gene is expressed in both cells, the sequence is yellow.

If the gene is expressed only in tumour cells, the sequence is red.

If the gene is expressed only in normal cells, the sequence is green.

The next step after hybridization is to generate an image using laser-induced fluorescent imaging.

The amount of fluorescence measured at each sequence specific location is directly proportional to the amount of mRNA with complementary sequence present in the sample and transformed into numbers will be the basis for the statistical analysis.

$I$ genes and $J$ hybridizations are collected into the $I \times J$ matrix $N$ with elements $n_{ij}$ - the gene expression level for each gene $G_i$ in hybridizations $H_j$.

Gene expression matrix needs to be preprocessed, for example the logarithm of the raw intensity values is taken or normalization of data is performed.
**DISCOVERY (CLUSTERING)**

Discovery of co-regulated groups of genes of 2 types of patients A and B.


Unsupervised machine learning method such as hierarchical clustering, k-means clustering or self-organizing maps.

Identification of the genes that are similarly expressed.

Detection of spatial or temporal expression patterns.

Dimension reduction of the gene expression matrix.
**CLASSIFICATION (CLASSIFICATION)**

Determine mathematical model well describing the classification rule used to distinguish the pre-defined classes.

Estimate the parameters of the mathematical function used in this model.

Estimate the accuracy of the predictor.

Class Prediction example: assignment of type to a new sample of gene expression matrix.

PATHWAY ANALYSIS

Biological interpretation of the list of genes selected in microarray analysis experiment.


Gene functional association networks for selected pathways.

Comparison of the gene expression levels of genes between groups of patients using such methodology as Student’s t-test, ANOVA, survival analysis, PCA, Correspondence Analysis.

**null hypothesis**
- given gene on the array is not differentially expressed between the two conditions under study

**alternative hypothesis**
- the expression level of that gene is different

**Student’s t-test:**
\[
\frac{\bar{x}_T - \bar{x}_N}{\sqrt{\frac{\text{var}_T}{n_T} - \frac{\text{var}_N}{n_N}}}
\]

50 genes identified at P<0.05
Are there significant?

**Distinction of Tumor vs. Normal 1000 genes.**

Source: Jon Pollack 2011: *Microarrays and Analysis of Hybridization Data*, Genomic Medicine
**CORRESPONDENCE ANALYSIS**

- **n gene profiles:** vectors in $m$-dimensional experiment space
- **m hybridisation profiles:** vectors in $n$-dimensional experiment space

Projection into a common subspace of low dimensionality for visualisation

Visualisation of hybridisations and genes at the same time

Reveals interdependencies (correspondence) between hybridisations and genes
The hybridizations are represented in \( n \)-dimensional gene space (here \( n=3 \)). The plane is selected such that the distance of hybridization vectors to the plane is minimal, thus conserving point-to-point distances among those vector points as well as possible.

Genes and tissues are typically classified using correlations of gross expression level. The net relationship between a pair of genes may be measured by partial correlation.
TABLES statement instructs PROC CORRESP to create a contingency table from raw, categorical data.

VAR statement instructs PROC CORRESP to read an existing contingency table.

BY statement separates analyses on observations in groups defined by the BY variables.

WEIGHT Statement specifies weights for each observation and indicates supplementary observations for simple correspondence analyses with VAR statement input.

ID statement (only with VAR statement) labels the rows of the tables with the ID values and places the ID variable in the output data set.

SUPPLEMENTARY statement specifies variables that are to be represented as points in the joint row and column space but that are not used in determining the locations of the other, active row and column points of the contingency table.
EXAMPLE: Alon et al. (1999): series of 62 Affimetrix GeneChip experiments upon normal (N) and cancerous (T) colon tissue.

The total $\chi^2$-statistic, which is a measure of the association between the rows and columns is 5008.79 and is explained equally for both the dimensions i.e. about 53.27% explain Dimension 1 and 46.73% explain Dimension 2. This indicates that the association between the row and column categories is essentially two dimensional.
The distant pair of T33 and T26 shows a low correlation of gene expression, although they are of the same tissue type.

Tissues T33 and n39 are located close together, although they are of different tissue types and from different individuals.

The normal cells are mostly distributed in the upper-right region, whereas the tumor cells are distributed in the lower-left region, so the visual separation is moderately good.
The distribution of the z-scores (i.e. difference divided by s.d.) between the mean expression level of a gene in tumor cells and that in normal cells.

Orange accession numbers are attached to the genes with z-scores larger than 3,
Blue accession numbers are attached to the genes with z-scores smaller than −3
MICROARRAYS DNA EXPERIMENTS:

- Enables the researchers to monitor the expression levels of thousands of genes simultaneously.
- Expression matrix can be used to detect the new subclasses of diseases, protect clinically important outcomes, such as the response to therapy and survival.
- Problem: large number of genes vs. relatively small number of experiments.

CORRESPONDENCE ANALYSIS:

- No parametrisation needed.
- Projection into a common subspace hybridisations and genes.
- Dimensionality reduction.
- The result of experiments can be used in medicine for comparing clinically relevant groups (e.g., healthy vs diseased).
THANK YOU